ChimpChunk2_2 Annotation

Lisa Sudmeier
Biology 4342
3/30/07

Introduction

Although DNA sequences can now be generated with high efficiency, the process of finding features such as genes and repetitive elements within these sequences is much slower and laborious. Fortunately, computational programs such as GENSCAN, Blast, and others have been developed to aid in the annotation process, making it less labor intensive. GENSCAN and other basic gene finders can predict features within sequences that are possible genes. These features can then be analyzed using programs like Blast to further investigate their function in the sequence (gene, non-gene, pseudogene, etc.).

Pre-Annotation

Before beginning to annotate ChimpChunk2_2, the repetitive elements were masked by using RepeatMasker on the entire sequence. The –nolow option was specified so that low complexity regions would not be masked. GENSCAN was then used to search for regions with possible intron-exon structure in this masked sequence, and four features were found. The first feature, in the region 23bp-7.8kb, was predicted to contain three exons encoding a 182 residue polypeptide. Further analysis suggested that this feature is a non-gene. The second feature predicted by GENSCAN was in the region 16.6kb-44.8kb and contained four exons thought to encode a 571 residue polypeptide. After investigation, it was determined that this feature encoded a region of the gene for the chimp ortholog of the human WDR22 protein (accession #: NP_003852). GENSCAN predicted a third feature with 3 exons in the 46.3kb-47.8kb region encoding a 412 residue polypeptide. This feature appears to be a pseudogene that arose from transposition of the DDX18 gene (accession #: NP_006764). The final feature predicted by GENSCAN was located from 53.99kb-116kb. It was predicted to have eight exons encoding a 351 residue polypeptide. Further analysis showed that this feature, like feature 2, encoded a region of the chimp ortholog for the human WDR22 protein (accession #: NP_003852).

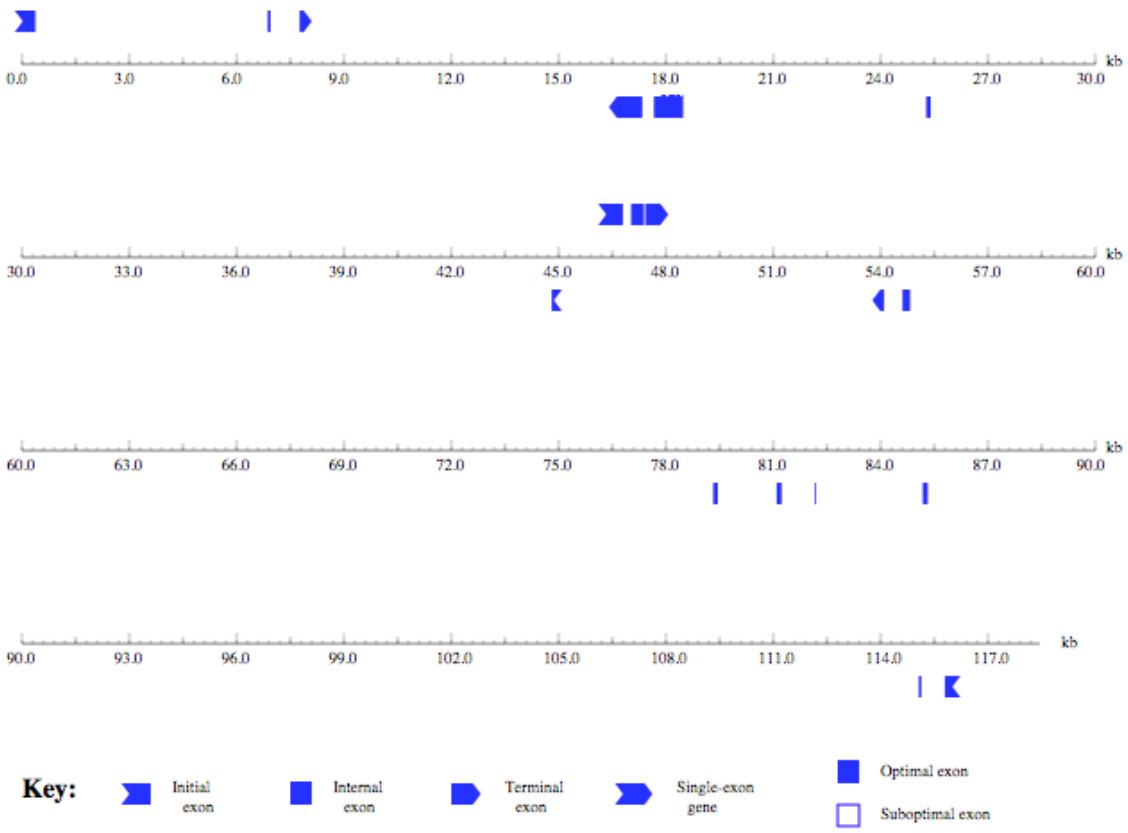
Feature	**Exons**	**Position**	**Characterization**
1	3	23bp-7.8kb	Non-gene
2	4	16.6kb-44.8kb	Gene (WDR22 ortholog)
3	3	46.3kb-47.8kb	Pseudogene (DDX18)
4	8	53.99kb-116kb	Gene (WDR22 ortholog)

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	23	376	354	1	0	53	53	201	0.083	8.58
1.02	Intr	+	6868	6948	81	0	0	-3	55	136	0.480	0.12
1.03	Term	+	7761	7874	114	2	0	88	48	121	0.522	5.69
1.04	PlyA	+	9276	9281	6							1.05
2.05	PlyA	-	12359	12354	6							1.05
2.04	Term	-	17339	16631	709	1	1	17	44	654	0.766	45.89
2.03	Intr	-	18481	17664	818	1	2	98	-58	1201	0.767	95.79
2.02	Intr	-	25400	25273	128	0	2	127	100	113	0.995	15.88
2.01	Init	-	44868	44808	61	0	1	67	106	26	0.226	3.76
2.00	Prom	-	45662	45623	40							-7.05
3.00	Prom	+	45968	46007	40							-10.45
3.01	Init	+	46331	46796	466	1	1	60	51	490	0.800	38.35
3.02	Intr	+	47026	47382	357	2	0	86	26	189	0.973	6.80
3.03	Term	+	47425	47840	416	2	2	-12	34	355	0.904	14.54
3.04	PlyA	+	52425	52430	6							1.05
4.09	PlyA	-	53512	53507	6							1.05
4.08	Term	-	54093	53992	102	0	0	73	48	72	0.981	-1.00
4.07	Intr	-	54825	54612	214	2	1	107	99	169	0.943	17.70
4.06	Intr	-	79450	79321	130	2	1	98	83	64	0.948	5.73
4.05	Intr	-	81233	81094	140	1	2	86	80	127	0.585	10.89
4.04	Intr	-	82186	82150	37	2	1	115	94	48	0.949	4.50
4.03	Intr	-	85320	85177	144	1	0	82	83	61	0.127	4.33
4.02	Intr	-	115140	115066	75	1	0	58	54	81	0.504	0.27
4.01	Init	-	116008	115795	214	1	1	84	99	187	0.455	16.40

Figure 1:
GENSCAN output of RepeatMasked ChimpChunk2_2 (text version)

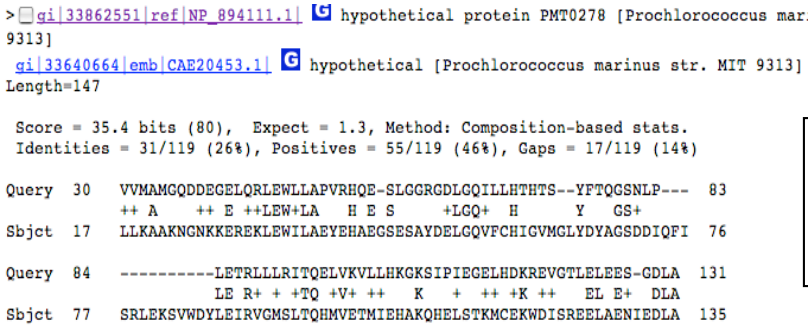
Figure 2:

GENSCAN predicted genes in sequence Pan



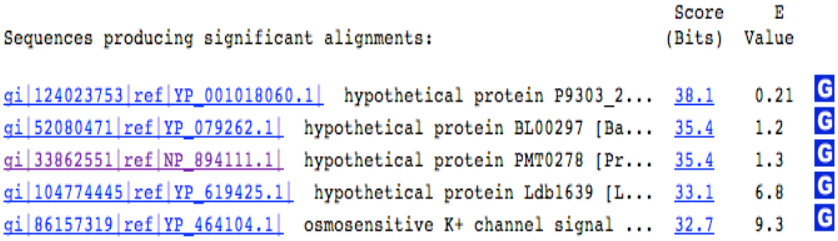
Analysis of Feature 1

Feature 1 was first analyzed using blastp to search the protein database with the amino acid sequence predicted by GENSCAN. Only five alignments were found (see Figure 4), none of which had a reliable E value (reliable is around 10^-10 or lower). There was only one alignment supported by experimental evidence, which had an E value of 1.3 (see Figure 3). Also, none of the alignments were for the same protein, which is suspicious because a real gene would only align with one protein or protein family. Futhermore, most of the alignments were with hypothetical proteins. Therefore, it was concluded that feature 1 is a non-gene, a feature that has characteristics of a gene, but doesn't resemble any known protein-coding sequences.



```
>gi|33862551|ref|NP_894111.1| hypothetical protein PMT0278 [Prochlorococcus marinus str. MIT
9313]
 gi|33640664|emb|CAE20453.1| hypothetical [Prochlorococcus marinus str. MIT 9313]
Length=147

 Score = 35.4 bits (80),  Expect = 1.3, Method: Composition-based stats.
 Identities = 31/119 (26%), Positives = 55/119 (46%), Gaps = 17/119 (14%)

Query  30   VVMAMGQDDEGELQRLEWLLAPVRHQE-SLGGRGDLGQILLHTHTS--YFTQGSNLP---  83
            ++ A    ++ E ++LEW+LA   H E S     +LGQ+  H      Y   GS+
Sbjct  17   LLKAAKNGNKKEREKLEWILAEYEHAEGSESAYDELGQVFCHIGVMGLYDYAGSDDIQFI  76

Query  84   ----------LETRLLLRITQELVKVLLHKGKSIPIEGELHDKREVGTLELEES-GDLA  131
                      LE R+ + +TQ +V+ ++    K    +  ++ +K ++   EL E+  DLA
Sbjct  77   SRLEKSVWDYLEIRVGMSLTQHMVETMIEHAKQHELSTKMCEKWDISREELAENIEDLA  135
```

Figure 3: A blastp alignment for feature 1



```
                                                                  Score     E
Sequences producing significant alignments:                       (Bits)  Value

gi|124023753|ref|YP_001018060.1|  hypothetical protein P9303_2...  38.1    0.21
gi|52080471|ref|YP_079262.1|  hypothetical protein BL00297 [Ba...  35.4    1.2
gi|33862551|ref|NP_894111.1|  hypothetical protein PMT0278 [Pr...  35.4    1.3
gi|104774445|ref|YP_619425.1|  hypothetical protein Ldb1639 [L...  33.1    6.8
gi|86157319|ref|YP_464104.1|  osmosensitive K+ channel signal ...  32.7    9.3
```

Figure 4: All of the blastp hits for feature1

Analysis of Feature 3

A search with the amino acid sequence predicted by GENSCAN and the protein database using blastp was also used to begin analysis of feature three. The output showed significant alignment of the DDX18 human protein with the amino acid sequence of feature 3 predicted by GENSCAN. Although the alignments had many gaps and mismatches, the E values were very low (much lower than 10^-10). The first alignment with experimental evidence had an E value of 1e-135 (see Figure 6). This suggested that feature 3 was not merely a non-gene, but either a gene or pseudogene. Blastn was then used to compare the GENSCAN predicted sequence for feature 3 with the whole ChimpChunk2_2 sequence to determine the intron/exon boundaries. Blat was then used to compare the sequence of these individual introns and exons with the human genome. All introns and exons had a high percent identity (98.8-100%) with a region on human chromosome 14. However, the human gene to which blastp had aligned this chimp amino acid sequence was located on chromosome 2. This suggests that a transposition of this gene, which is actually located on chromosome 2, occurred before human evolved

from chimp resulting in a pseudogene on chromosome 14 in both species. The likelihood of this feature encoding a pseudogene was further supported by the Blat results showing that the percent identity between the chimp sequence and human genome was relatively constant for both introns and exons. Pseudogenes are often marked by similarity in exon and intron divergence because the exons are not under more selective pressure than the introns, as is the situation with real genes. Therefore, feature 3 is very likely a pseudogene.

```
                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

gi|114580554|ref|XP_515753.2|  PREDICTED: DEAD (Asp-Glu-Ala-As...   492    2e-137
gi|12654791|gb|AAH01238.1|  DEAD (Asp-Glu-Ala-Asp) box polypep...   486    1e-135
gi|38327634|ref|NP_006764.3|  DEAD (Asp-Glu-Ala-Asp) box polyp...   486    1e-135
gi|7022744|dbj|BAA91709.1|  unnamed protein product [Homo sapiens   482    1e-134
gi|26344732|dbj|BAC36015.1|  unnamed protein product [Mus musculu   474    5e-132
gi|31981163|ref|NP_080136.2|  DEAD (Asp-Glu-Ala-Asp) box polyp...   473    7e-132
gi|12860207|dbj|BAB31877.1|  unnamed protein product [Mus musculu   470    6e-131

>gi|38327634|ref|NP_006764.3|  DEAD (Asp-Glu-Ala-Asp) box polypeptide 18 [Homo sapiens]
 gi|20532388|sp|Q9NVP1|DDX18_HUMAN  ATP-dependent RNA helicase DDX18 (DEAD box protein 18) (Myc-regulated
DEAD box protein) (MrDb)
 gi|119615599|gb|EAW95193.1|  DEAD (Asp-Glu-Ala-Asp) box polypeptide 18, isoform CRA_b [Homo
sapiens]
 gi|119615600|gb|EAW95194.1|  DEAD (Asp-Glu-Ala-Asp) box polypeptide 18, isoform CRA_b [Homo
sapiens]
Length=670

 Score =  486 bits (1250),  Expect = 1e-135, Method: Composition-based stats.
 Identities = 297/509 (58%), Positives = 324/509 (63%), Gaps = 98/509 (19%)

Query  1    MSHLLMKLLRKKIKKWNLKLRQWNLKLQGASNLTLSE----------------------  37
            MSHL MKLLRKKI+K NLKLRQ NLK QGASNLTLSE
Sbjct  1    MSHLPMKLLRKKIEKRNLKLRQRNLKFQGASNLTLSETQNGDVSEETMGSRKVKKSKQKP  60

Query  38   -------TQNTDVSEETTGGGKVKKSKHSMNV---GLSDAQNG-----------------  70
                  TQN  +S+E  G  KV KS     V   G +  Q+
Sbjct  61   MNVGLSETQNGGMSQEAVGNIKVTKSPQKSTVLTNGEAAMQSSNSESKKKKKKKRKMVND  120

Query  71   ---DVSQEAVENIKVKKSPQKSTVLTNGEAAMQSPNSESKKKK-----------------  110
              D  +   EN K K   + +       E  ++ P+++   + +
Sbjct  121  AEPDTKKAKTEN-KGKSEEESAETTKETENNVEKPDNDEDESEVPSLPLGLTGAFEDTSF  179

Query  111  RKMVNDAESDTKKAKTE----NKGESEEESAKSPKETENNVEKPDDEDD---------TE  157
               + N    +T KA  E    N  E + +S +   E  + +                E
Sbjct  180  ASLCNLVNENTLKAIKEMGFTNMTEIQHKSIRPLLEGRDLLAAAKTGSGKTLAFLIPAVE  239

Query  158  LIVKLNFMPRNGTGVLILSPTRELDMQTFGVLKELMSHHVHTYGLIMGGSNRSAEAQKLA  217
            LIVKL FMPRNGTGVLILSPTREL MQTFGVLKELM+HHVHTYGLIMGGSNRSAEAQKL
Sbjct  240  LIVKLRFMPRNGTGVLILSPTRELAMQTFGVLKELMTHHVHTYGLIMGGSNRSAEAQKLG  299

Query  218  NGINITVVTPGCLLDHMQNIPGFMYKNLQCLVIDEADRILDVEFEEELKQIIKLLPT---  274
            NGINI V TPG LLDHMQN PGFMYKNLQCLVIDEADRILDV FEEELKQIIKLLPT
Sbjct  300  NGINIIVATPGRLLDHMQNTPGFMYKNLQCLVIDEADRILDVGFEEELKQIIKLLPTRRQ  359

Query  275  -----------LEGPARTSLKKEPLYVGVDDDKANATVDGLEQGYVVCPSEKRFLLLFTF  323
                      +E  AR SLKKEPLYVGVDDDKANATVDGLEQGYVVCPSEKRFLLLFTF
Sbjct  360  TMLFSATQTRKVEDLARISLKKEPLYVGVDDDKANATVDGLEQGYVVCPSEKRFLLLFTF  419

Query  324  LKKNQKKKLMAFFSSCMSVKYHYELLNYIDLPILAIHGKQKQNKHTTTFFQFCNADSGTL  383
            LKKN+KKKLM FFSSCMSVKYHYELLNYIDLP+LAIHGKQKQNK TTTFFQFCNADSGTL
Sbjct  420  LKKNRKKKLMVFFSSCMSVKYHYELLNYIDLPVLAIHGKQKQNKRTTTFFQFCNADSGTL  479

Query  384  LCTDVAARELDITEVNWIVQYDPPDDPKE  412
            LCTDVAAR LDI EV+WIVQYDPPDDPKE
Sbjct  480  LCTDVAARGLDIPEVDWIVQYDPPDDPKE  508
```

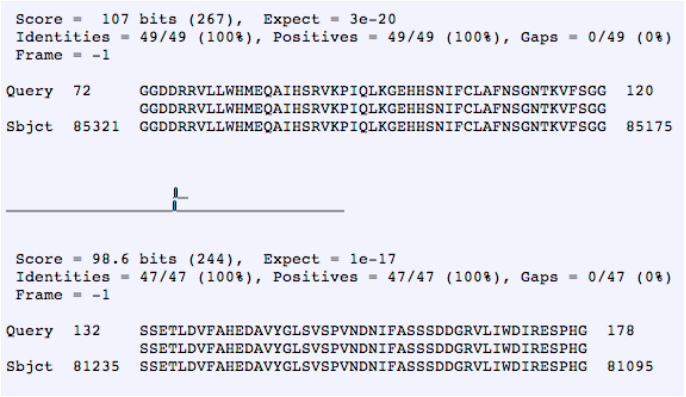
Figure 5: First 7 blastp hits for feature 3

Figure 6: Blastp alignment for feature 3 with human DDX18

Analysis of Features 2 and 4

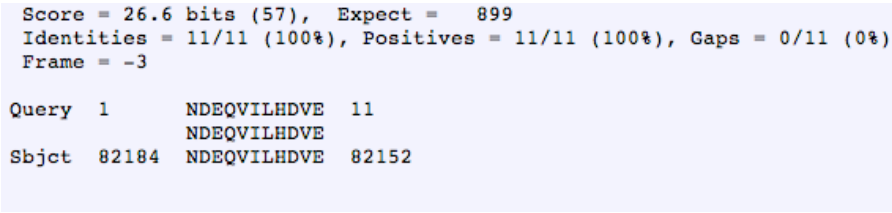
When blastp was used to search with the amino acid sequences of features 2 and 4, each aligned with different regions of the same protein, human WDR22. Feature 4 matched the first part (residues 1-294), and feature 2 aligned with amino acids 306-631. These alignments were of very high quality with low E values. This suggests that these

features could either be part of a gene or a pseudogene. For this region to be the ortholog of human WDR22, the whole peptide sequence of the protein would have to be encoded in the ChimpChunk. Therefore, the peptide sequence of the human protein was obtained from Ensembl and a tblastn was performed to align the peptide sequence of this protein with the entire sequence of the chimpchunk. The output showed that chimp sequence aligned with the entire protein except for the absence of 11 amino acids (see Figure 7). To determine whether these amino acids were actually missing in the chimp sequence or just missed by the Blast program, another tblastn was performed with the missing 11 amino acids and the whole ChimpChunk. However, before performing this alignment, the expect value was changed from 10 to 100,000 to ensure that such a short sequence, if present, would align. The output showed that the 11 amino acids were present in the ChimpChunk (Figure 8). The initial Blat alignment did not find this short region of amino acids because the chimp sequence contains an extra intron between these 11 amino acids and the next exon. The human sequence includes these amino acids in the neighboring exon.



```
 Score =  107 bits (267),  Expect = 3e-20
 Identities = 49/49 (100%), Positives = 49/49 (100%), Gaps = 0/49 (0%)
 Frame = -1

Query  72     GGDDRRVLLWHMEQAIHSRVKPIQLKGEHHSNIFCLAFNSGNTKVFSGG  120
              GGDDRRVLLWHMEQAIHSRVKPIQLKGEHHSNIFCLAFNSGNTKVFSGG
Sbjct  85321  GGDDRRVLLWHMEQAIHSRVKPIQLKGEHHSNIFCLAFNSGNTKVFSGG  85175


 Score = 98.6 bits (244),  Expect = 1e-17
 Identities = 47/47 (100%), Positives = 47/47 (100%), Gaps = 0/47 (0%)
 Frame = -1

Query  132    SSETLDVFAHEDAVYGLSVSPVNDNIFASSSDDGRVLIWDIRESPHG  178
              SSETLDVFAHEDAVYGLSVSPVNDNIFASSSDDGRVLIWDIRESPHG
Sbjct  81235  SSETLDVFAHEDAVYGLSVSPVNDNIFASSSDDGRVLIWDIRESPHG  81095
```

Figure 7: tblastn of WDR22 peptide sequence to ChimpChunk2_2 (the "missing 11 amino acids are between residues 120 and 132)



```
 Score = 26.6 bits (57),  Expect =   899
 Identities = 11/11 (100%), Positives = 11/11 (100%), Gaps = 0/11 (0%)
 Frame = -3

Query  1      NDEQVILHDVE  11
              NDEQVILHDVE
Sbjct  82184  NDEQVILHDVE  82152
```

Figure 8: tblastn of "missing" 11 amino acids to ChimpChunk2_2 sequence

The alignment of the entire WDR22 peptide sequence with the ChimpChunk2_2 sequence suggested that features 2 and 4 were a gene, but some questionable areas of the alignment prevented this conclusion from being made too soon. The tblastn output comparing the human peptide sequence to the ChimpChunk showed a gap in the ChimpChunk sequence between bases 17546 and 17665 (Figure 9). According to the tblastn results, this gap was right in the middle of an exon. Although this gap could be showing a difference between the human and chimp orthologs of this protein, it would be unlikely that two organisms so closely related would have orthologs that differed by 40 amino acids in a row. To investigate the cause of this gap, the region of the ChimpChunk corresponding to the gap was extracted. The extracted sequence showed that this region of the chimp genome hadn't been sequenced, shown by bases represented with 'N'

(Figure 10). No further analysis of the gap was needed because the absence of sequence would, itself, result in a gap.

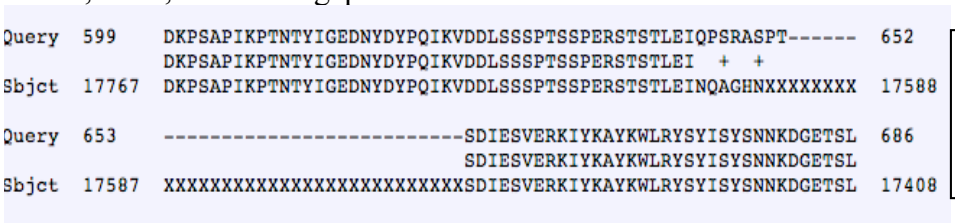


Figure 9: Gap in tblastn alignmen

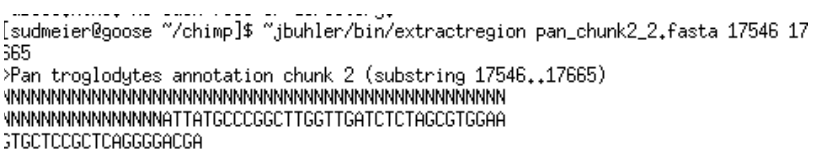


Figure 10: Gap is result of missing sequence

The last region to be investigated before there was enough confidence to call this a gene was an extra sequence in the ChimpChunk that appeared in the initial blastp output of feature 4 (Figure 11). To determine where this extra sequence was located in respect to the peptide sequence, the amino acids flanking the extra sequence were mapped to the end and beginning of two neighboring exons in the protein (Figure 12). Therefore, this extra sequence appears to come from the intron between those two neighboring exons. Two possible explanations for this extra sequence are that 1) an extra exon is present in the chimp gene or 2) GENSCAN falsely predicted this as protein coding sequence, but it is actually just a part of the chip's intron that GENSCAN used to combine two exons. A tblastn analysis comparing this extra amino acid sequence to the human genome showed that the sequence is present in the human intron between the two neighboring exons (in the region 50,588,828 - 50,588,685bp). Further investigation into the cause of this extra sequence would involve wet lab experiments to determine whether the extra sequence is included in the expressed chimp WDR22 ortholog.

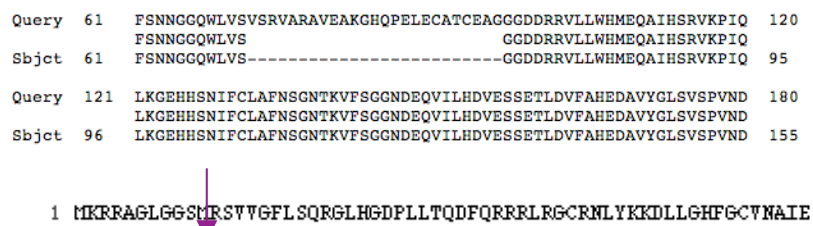


Figure 11: Extra sequence included in chimp exon

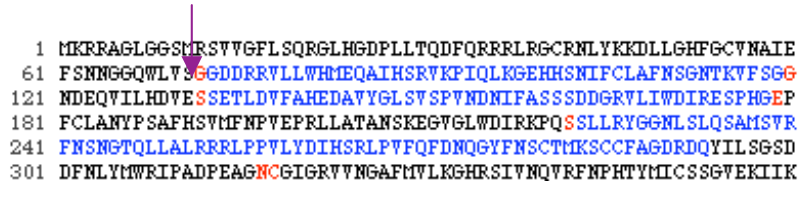


Figure 12: Arrow points to exon boundary where extra sequence was predicted in chimp

The presence of sequence to code for all regions of the human WDR22 protein in the ChimpChunk as well as the high quality alignment strongly support the conclusion that features 2 and 4 are a gene encoding the chimp ortholog of WDR22. Final evidence for this conclusion was obtained by using Blat to compare the human WDR22 peptide sequence with the human genome. The only match was the WDR22 protein, itself, showing that WDR22 is not part of a protein family. This result eliminates the possibility

that features 2 and 4 of the ChimpChunk may code for another protein in the same family as WDR22 rather than the chimp ortholog of this protein. Lastly, the human WDR22 peptide sequence was compared to the chimp genome using Blat. Two hits resulted from this search, one with 99.8% identity and the other with 90.5% identity. An identity greater than 99% is expected for orthologs, supporting that conclusion that this ChimpChunk contains the gene encoding the ortholog of the human WDR22 protein. The other hit from this Blat search is likely another protein of similar structure.

Repeats

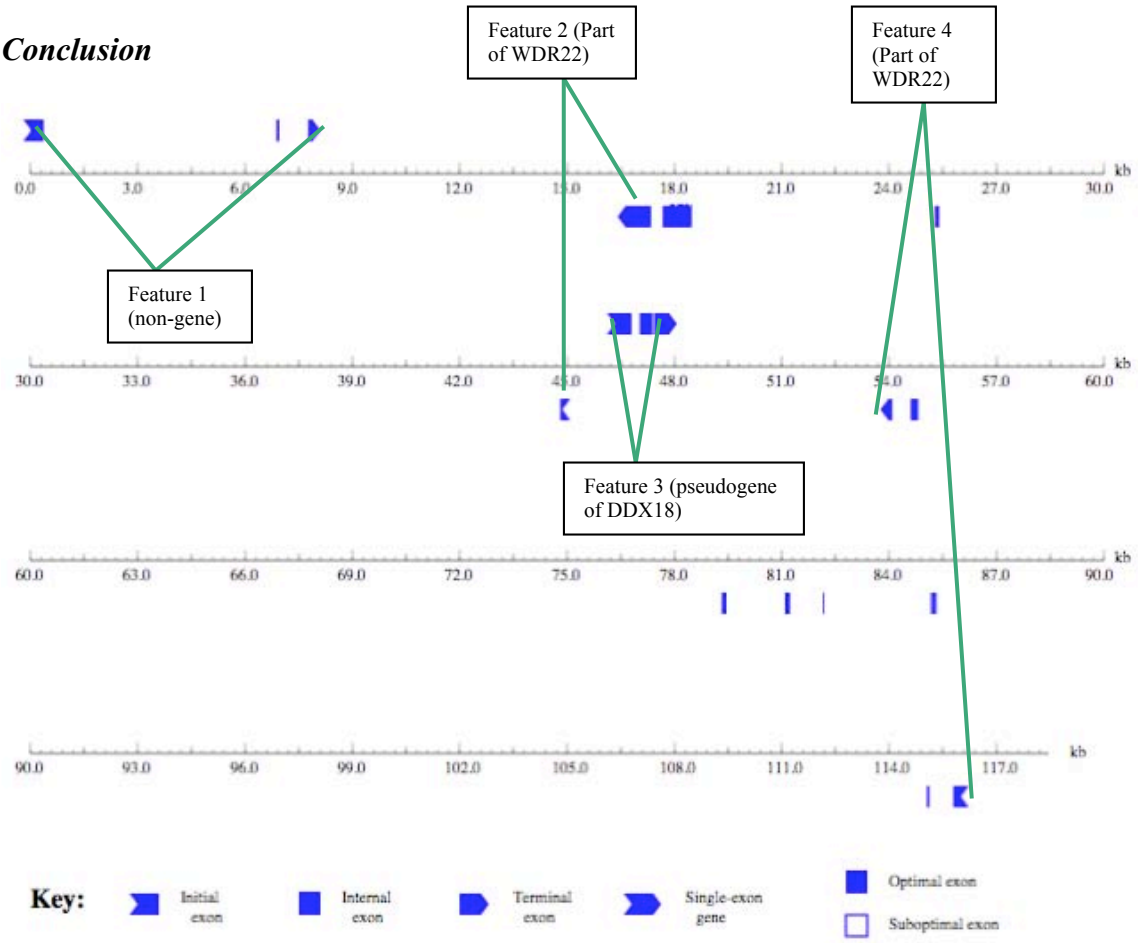
```
==================================================
file name: pan_chunk2_2.fasta
sequences:          1
total length:     118450 bp  (116915 bp excl N-runs)
GC level:         41.62 %
bases masked:      41464 bp ( 35.01 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:               99        21417 bp    18.08 %
      ALUs           62        16262 bp    13.73 %
      MIRs           37         5155 bp     4.35 %

LINEs:               37        13753 bp    11.61 %
      LINE1          21        10179 bp     8.59 %
      LINE2          14         2921 bp     2.47 %
      L3/CR1          2          653 bp     0.55 %

LTR elements:         4         1042 bp     0.88 %
      MaLRs           3          482 bp     0.41 %
      ERVL            1          560 bp     0.47 %
      ERV_classI      0            0 bp     0.00 %
      ERV_classII     0            0 bp     0.00 %

DNA elements:        20         5140 bp     4.34 %
      MER1_type      15         3162 bp     2.67 %
      MER2_type       3         1860 bp     1.57 %

Unclassified:         0            0 bp     0.00 %

Total interspersed repeats:    41352 bp    34.91 %


Small RNA:            2          136 bp     0.11 %

Satellites:           0            0 bp     0.00 %
Simple repeats:       0            0 bp     0.00 %
Low complexity:       0            0 bp     0.00 %
```

Figure 13 (left): Full summary table of repeats

Figure 14 (below): Repeats over 500 bp

Begin Pos.	End Pos.	Repeat Class/Family	Element Size
98124	99122	DNA/MER1_type	999
72336	73240	LINE/L1	905
38806	39694	LINE/L1	889
89972	90814	LINE/L1	843
65359	66059	DNA/MER2_type	701
60876	61543	LINE/L2	668
66605	67225	LINE/L1	621
76966	77534	LINE/L1	569
68380	68939	LTR/ERVL	560
89444	89970	LINE/L1	527
64556	65058	DNA/MER2_type	503

Conclusion


Feature 2 (Part of WDR22)
Feature 4 (Part of WDR22)
Feature 1 (non-gene)
Feature 3 (pseudogene of DDX18)
kb
0.0
3.0
6.0
9.0
12.0
15.0
18.0
21.0
24.0
27.0
30.0
33.0
36.0
39.0
42.0
45.0
48.0
51.0
54.0
57.0
60.0
63.0
66.0
69.0
72.0
75.0
78.0
81.0
84.0
87.0
90.0
93.0
96.0
99.0
102.0
105.0
108.0
111.0
114.0
117.0
Key:
Initial exon
Internal exon
Terminal exon
Single-exon gene
Optimal exon
Suboptimal exon